Further exploration of the possibilities and pitfalls of multidimensional scaling as a tool for the evaluation of the quality of synthesized speech

Anna C. Janska and Robert A. J. Clark. Further exploration of the possibilities and pitfalls of multidimensional scaling as a tool for the evaluation of the quality of synthesized speech. In The 7th ISCA Tutorial and Research Workshop on Speech Synthesis, pages 142-147, 2010.

Further exploration of the possibilities and pitfalls of multidimensional scaling as a tool for the evaluation of the quality of synthesized speech

Anna C. Janska[1], *Robert A. J. Clark*[2]

[1]IMPRS NeuroCom, University of Leipzig, Germany
[2]CSTR, The University of Edinburgh, U.K.

janska@rz.uni-leipzig.de, robert@cstr.ed.ac.uk

Abstract

Multidimensional scaling (MDS) has been suggested as a useful tool for the evaluation of the quality of synthesized speech. However, it has not yet been extensively tested for its application in this specific area of evaluation. In a series of experiments based on data from the Blizzard Challenge 2008 the relations between Weighted Euclidean Distance Scaling and Simple Euclidean Distance Scaling is investigated to understand how aggregating data affects the MDS configuration. These results are compared to those collected as mean opinion scores (MOS). The ranks correspond, and MOS can be predicted from an object's space in the MDS generated stimulus space. The big advantage of MDS over MOS is its diagnostic value; dimensions along which stimuli vary are not correlated, as is the case in modular evaluation using MOS. Finally, it will be attempted to generalize from the MDS representations of the thoroughly tested subset to the aggregated data of the larger-scale Blizzard Challenge.

1. Introduction

Evaluation plays a vital role in advancing the state-of-the-art in text-to-speech (TTS) synthesis research. However, especially for the subjective evaluation of the quality of synthesized speech this is far from trivial [14]: the output generated cannot simply be measured according to its accuracy, as there is no such thing as "right speech", or "wrong speech". To this day, it is not entirely known which dimensions listeners consider and to what extent they do so when evaluating the quality of synthesized speech.

Mean opinion scores (MOS) are a common method for gathering subjective judgments of speech quality. MOS are derived by analysis of untrained listeners' ratings of stimuli along a scale. These scores are only valid within the context they were tested in, as one sample whose quality differs drastically from that of the samples surrounding it is likely to be appointed more extreme scores than it would be within samples of similar quality. Furthermore, listeners differ, and these values are not absolute [7, p. 2167]. As such, [16] suggest to consider MOS tests as "ranking tests". (p. 537)

Also, since evaluation of speech quality is based on abstract categories, identifying distinct dimensions such as segmental quality or appropriateness of intonation, can be hard for experimental participants. Furthermore, when asked to listen to one of these dimensions, more perceptually salient dimensions tend to influence their judgments [12, 16]. A study conducted by [1] found that judgments across dimensions were highly correlated, which, they suggested, indicates that listeners were not assessing clearly distinct dimensions of the systems under investigation. [8] and [17] have reported similar results.

Hence, reliable results will only be obtained, when speech quality is tested as a whole, and listeners are not asked to make distinctions on a level lower than that. As such,

> [t]he single composite judgment of quality provided by MOS testing is essential for acceptance testing, but it does not tell us *why* the quality is good or bad [7, p. 2168].

Currently, one of the biggest issues in subjective evaluation of speech quality is not ranking stimuli or systems, but explaining *how* these ranks were derived. Up to date, no system will be mistaken for a human speaker in all its output; systems are far from sounding perfectly natural, and various imperfections are aggregated. This is less trivial than it sounds; listeners have a good notion of a system's speech quality, and a system that does not sound right is easily detected. It is much harder, though, to explain why it does not sound right. Ultimately, an objective measure for TTS evaluation is the desirable research goal, but this will not be feasible until we have gained better understanding of listeners' perceptual behaviour in evaluating synthesized speech; in particular this means: knowing how different dimensions are weighted in human perceptual processing of synthesized speech.

A few years ago, multidimensional scaling (MDS) has been discovered as a suitable tool for understanding "what acoustic cues listeners attend to by default when asked to evaluate synthetic speech" [12, p. 1].

This article addresses several questions that are vital in determining whether the introduction of MDS as a standard tool for large scale evaluation of the quality of synthesized speech is feasible, and if so, whether its results yield more insights than the already established MOS. Since similarity-difference judgements, as they are required by MDS are fairly costly to gather, it is vital for us to investigate if aggregation of data across listeners and/or across stimuli of a system affect the results. Furthermore, it needs to be investigated whether MDS arrives at compensating for the shortcomings of MOS.

2. Background

MDS can be used as a psychological model that transforms judgments of similarity (e.g. "Are these two stimuli the same, or are they different?") into metric distances.

> The most common approach is to hypothesize that a person, when asked about the dissimilarity of pairs of objects, acts as *if* he or she computes a distance in his or her "psychological space" of these objects. [2, p. 11]

[13] refers to MDS as "most appropriate when the goal of your analysis is to find the structure in a set of distance measures between objects or cases" (p.13). This is indeed what we aspire in the evaluation of speech quality: we want to understand the structure of the components that contribute to what we then perceive as the degree of naturalness of synthesized speech.

Multidimensional scaling maps "proximities p_{ij} [...] into distances of an m-dimensional MDS configuration X [..., defined] by a *representation function* $f(p_{ij})$ that specifies how the proximities should be related to distances $d_{ij}(X)$." [2, p. 37]

MDS is very attractive tool for evaluation owing to its robustness as a measure, as it is not susceptible to non-systematic missing data, not confined to a certain level of measure, and distribution-free. The graphical output is easily interpretable, and allows to explain the underlying structure in data [5] by identifying points distant from one another, of which some qualities are already known; based on the prior knowledge of these characteristics, a substantive criterion that could have induced experimental participants to distinguish between these objects, i.e a criterion that could have led them to place the stimuli at opposite ends of a dimensions, is determined [2, p. 11].

2.1. Applying MDS in large scale evaluation

For the reasons named above, MDS was considered for large-scale evaluation of TTS synthesis. The Blizzard Challenge, which has been held annually since 2005, is a "research exercise" [11] for TTS systems, and has in recent years also gathered data suitable of MDS analysis, as well as MOS of overall naturalness, and MOS rating the similarity of test sentences to reference sentences [6]. Due to the large amount of data entered in the Blizzard Challenge it is not feasible that one experimental participant is presented with all stimuli that are used for evaluation, as the quantity is sheer overwhelming. Listeners are appointed to groups, and each group is presented with different sets of sentences for a subset of the entries into the competition. Thus the data gathered are aggregated across listeners and across sentences for each system.

[14] conducted the first large-scale evaluation on these data using MDS and succeeded in replicating the findings of [12]'s pilot study. This more extensive study is the first one to include listeners from different vocational backgrounds and thus the first study to provide conclusive evidence supporting the claim that MDS is a valid means for the evaluation of speech quality of synthesized speech. What is still problematic, though, is that claims about perceptual dimensions were made, ignoring the weights individual listeners appoint to these. The open question really is whether aggregated data is representative of *any* listener at all, or whether it creates an artifact. The *Problem of Aggregation*, i.e. the representation of individuals' biases in the data is one of the biggest problems in data analysis of such a large scope.

It is possible to account for individual biases in MDS: Weighted Euclidean MDS is based on the assumption that individuals and groups may have some distinct perspectives, while they still share some common features with others. The Problem of Aggregation is tackled by presuming a *Group Space* X_{ij}, which consists of a fixed set of dimensions, and a *Subject Space*, in which the dimensions constituting the Group Space are appointed a weight between 0 and 1. ($[0...w_{ia}...1]$).

The weights can be interpreted as *importance* or *salience* of a dimension, and according to the pattern of these (i.e. their relative importance), a subject can be described in the form of their individual Subject Space. However, the major disadvantage of this approach is that a full data matrix is required, i.e. *no* aggregation of data is possible. Thus large-scale evaluation is not feasible at all, since every single point in the object space needs to be compared with all other points and by every participant.

To investigate whether MDS on aggregated data across participants as well as across stimuli yield results that are representative of the individual listeners, an experiment was conducted on a subset of the Blizzard Data 08.

3. Method

Thus comparisons were made

- between Euclidian Distance Scaling (EDS) and Weighted Euclidean Distance Scaling (WEDS) on a full data matrix to investigate whether aggregation of listener judgements significantly affects the MDS representation.
- between the ranks derived from MOS, EDS, and WEDS to determine how these less tried-and-tested means of evaluation of speech quality compare to the already established method of gathering MOS.
- between MOS and the position of stimuli in the object space generated by MDS.
- between the ranks of system averaged across sentences of the subset and across whole Blizzard test set to investigate how representative the subset is of the entire set.

3.1. Stimuli

A subset of entries in the Blizzard challenge 2008, test set A were selected to conduct more extensive MDS analysis. These stimuli consisted of pairs of stimuli from 5 different system entries. 4 of these systems were TTS systems and one was the natural speech control. In an attempt to obtain stimuli that stretch across the entire range of quality of our stimulus space, a ball-point estimate of the best and the worst stimulus of 5 systems was made: All comparisons of several groups were ranked according to their distance from the natural stimulus. The best 4 systems were selected, and then for each of these systems the stimulus receiving the most "different" ratings in comparison with a natural stimulus was selected. It goes without saying that these estimates cannot serve as legitimate manner of ranking, as the stimuli varied as well as the listeners, but simply an attempt to select stimuli of a range of qualities for each system. Natural speech recordings were included to "anchor the scale" in the MOS task, in which they should ideally be given the perfect score. [16, p. 537]. The length of the sentences ranges from 1.38 to 3.4 seconds, and from 8 to 15 syllables.[1] (cf. Figure 1)

3.2. Procedure

Altogether 30 participants from different vocational backgrounds were tested, who were were native speakers of some variety of English. The pool of participants was self-selecting: Participants were chosen on a first-come, first-serve basis in their response to an advertisement. They were paid 7 Pounds Sterling to take the experiment, which took none of the participants longer than 40 minutes to complete. The experiment was conducted in a computer lab. Instructions, as well as stimuli were represented on the 20 inch screen of an iMac computer as a

[1]The stimuli used in the experiment can be accessed at http//:homepages.inf.ed.ac.uk/s0674876/listening_test_july_2009_wavfiles/

label	type	sentence	syllables	duration
T1	natural	For good measure, he offered an unreserved apology.	15	2.9s
T2	synthesized	Billy could help Saxon little in her trouble.	12	2.2s
T3	synthesized	UCA based air traffic controllers are also unsettled.	15	3.4s
T4	synthesized	We are pulling on in the morning to circle city.	14	2.2s
T5	synthesized	I believe the two years suspension are harsh.	11	2.4s
B1	natural	Power cuts affect refrigerated medicines and food stuffs.	15	3.4s
B2	synthesized	But they can live in a pigsty.	8	1.38s
B3	synthesized	He was puzzled by the slowness of its progress.	12	2.2s
B4	synthesized	Thus he waited, keeping perfectly quiet.	11	2.4s
B5	synthesized	The bloodshed was not confined to Copenhagen.	12	2.5s

Figure 1: Stimulus sentences

web page within a Firefox browser window on full screen mode. Answers were given by clicking the respective radio-button on the screen, using an optical mouse. The subjects listened to the stimuli with closed-back Senheiser headphones and at a volume level they could adjust themselves.

3.2.1. Part 1 – Multidimensional Scaling (MDS) design

To facilitate producing a coordinate space of the stimuli in terms of how similarly they are perceived by each group of listeners through a multidimensional scaling protocol, each stimulus was paired with every other stimulus, so that paired comparisons between all stimuli were made, in both orders. This was done because the order of presentation within a stimulus pair could also affect subjects' perception of similarity and difference between sentences. The stimulus pairs were presented in random order. Participants had to decide whether both items of the pair were equal or different in their degree of naturalness. As in [12]'s pilot study, listeners

> were not instructed to listen to any one acoustic characteristic of the stimuli, or to any specific psychoacoustic construct (e.g., listening effort, pleasantness, pronunciation etc) such as have been used in previous evaluation studies e.g., [[15]]. The task was simply to make a simple binary decision about the degree of similarity in naturalness of each pair of stimuli.

3.2.2. Part 2 – Mean opinion scores (MOS)

Part 2 was devised to rank the systems according to their naturalness. The listeners' scores were conditional similarity data, which means that values cannot be compared directly between subjects [4, p. 14], but they provided a valid basis for generating ranks for the systems. Each stimulus was presented three times in random order. Participants had to rate on a scale from 1 to 10 (1 being the lowest, and 10 the highest), how natural a sentence sounded.[2] Part 2 of the experiment was presented after part one to ensure that subjects were familiar with the overall variation in quality of the stimuli by this point, to try to encourage a wider use of the scale provided.

4. Results

The data of part 1's 30 participants, resulting in 300 cases for 10 objects, which are 3000 edges, out of which 25 were missing, was put into a full distance matrix. The proximities are stacked in 10x10 matrices across columns. *similar* judgments were coded as 0, *different* judgments as 1. The experimental results were analyzed with PASW Statistics 17.0. MDS graphs were generated with the PROXSCAL function, which includes the *Identity Euclidean* function, and the *Weighted Euclidean Distance* function.

4.1. Comparing Weighted to Simple Euclidean MDS

Weighted Euclidean MDS was performed in two dimensions on an ordinal level, untieing ties, applying transformations to each point individually. Stress-1 is 0.14, and Dispersion accounted for (D.A.F.) is 0.98, which is a reasonable fit. So for now we will limit ourselves to two dimensions in favour of ease of interpretation of the graphical representation of the stimulus space. Analysis is done visually and auditively.

On a first glance, it is visible that the perceptual space can be divided into three groups (cf. Figure 2):

1. The two natural recordings, T1 and B1, are clustered together clearly distinct from the other stimuli. Hence we can deduce that experimental participants perceived a clear distance between those and the synthesized stimuli. This supports [9]'s claim that "even the best examples of speech from TTS systems are unlikely to be mistaken for natural speech". (p. 107)
2. There is a central section of the space, constituted by B4, B3, T5, T2 and T4, in which the largest number of systems are located.
3. There is a group of systems lagging behind, being T3, B2 and B5.

We attempt to further analyze the space by organizing the stimuli into clusters according to their auditory features. Obviously, the two natural recordings build one such cluster. B4, T5 and B3 all are characterized by good prosody, as the intonation is vivid and not flat. Thus they can be clustered together. However, B3 has some problems with joins, so that while the overall intonation is good, there are little "jumps in pitch" in between. Problems of joins can also be identified in B2, B5 and T3, so these are clustered together. T4, T2, B2, and B5 build a cluster of stimuli with bad intonation. B5 resembles an utterance of a NNS who speaks English rather well, while transferring their own language's "sentence melody" into English. B2 has rather flat intonation and the final segment in *pigsty* sounds somewhat clipped. T5 is unique in that it has a sort of echoing quality. The output of Weighted Euclidean Distance Scaling organizes the output along fixed axes, based on the salience appointed to them by listeners. The image in the middle of figure 2 shows these weights. The angles of the subject vectors represent differences between subjects. We can see that listeners do *not* fall into two clearly distinct groups, so an average across listeners will not create an artifact. Based on the clusters identified above, we can determine the parameters which change along the two axes.

[2] Generally, in MOS tasks, measures between 1 and 5 are used [9, 7, 10]. However, for this experiment, a larger range was chosen to try to encourage a wider dispersal of ratings, in the hope that this would generate bigger gaps between the systems' ratings and that distinct ranks could be clearly established.

Along dimension 1 (d1) the goodness of joins changes, improving as we move to the right side of the space. Along the vertical axis, dimension 2 (d2) varies in regards to goodness of intonation, being the most natural in the higher and the least natural in the lower regions of the space.

Overall naturalness is visible as the distance between T1 or B1 and the respective stimuli.

Having been able to create a stimulus space that is organized along the axes of perceptual dimensions from similarity-difference judgments substantiates claims already made by [7, 12, 14]: MDS allows to organize synthesized speech stimuli according to their naturalness, on the sole basis of data generated from similarity-difference judgments. However, so far, we have only employed Weighted MDS, like [7]. We will now investigate whether considering listeners' individual weights of dimensions is in fact absolutely necessary, or whether Simple Euclidean MDS generates output comparable to that of Weighted Euclidean MDS.

For that purpose, a Simple Euclidean MDS was conducted on an ordinal level, untieing ties. Stress 1 is 0.15, which is an acceptable fit, and D.A.F is 0.98. Dimensions are not predefined by the output of the graphical representation. Since the natural stimuli should be perfect in all dimensions, the space can be mirrored and/or rotated so that they are located at the extremes of both axes. We rotated the output graph and flipped it vertically to resemble that of Weighted Euclidean MDS (cf. Figure 2). Now it is possible to mark the dimensions of noisiness of speech signal and quality of prosody as done previously.

The only difference between the two representations is that in Simple Euclidean MDS B2 is somewhat lower and T2 is slightly higher up on d2 than it is in the other graph. From this we can conclude that the tested group was sufficiently homogeneous to generate comparable graph outputs for Simple and Weighted Euclidean MDS. This supports [12]'s assumption that the MDS representation they generated actually is representative of an average listener and not just an artifact resulting from the interference of different groups' perceptual patterns.

This finding is indeed good news for large-scale evaluation, as it provides support for the assumption that aggregating data across listeners is indeed permissible.

4.2. Comparing MOS to MDS representations

Correlations between dimensions and MOS were tested: For the Weighted MDS, there were significant positive correlations between d1 and MOS, r=.802, p(two-tailed) < 0.01, as well as between d2 and MOS, r=.847, p(two-tailed) < 0.01. There is no significant correlation between d1 and d2, which indicates that in our analysis of the stimulus space above we have indeed identified two discrete factors that influence listeners' judgments of speech quality. This justifies the larger cost of collecting MDS data as compared to MOS, which yield correlated scores in modular evaluation.

According to [7], the MOS, as were collected in part 2 should have correlates in the output of part 1. Ranks were computed for two-dimensional MDS representations of Weighted Euclidean distance measures with transformations applied to weights individually as well as with weights transformed simultaneously, MDS representations of Simple Euclidean distance measures, as well as for MOS [3] (cf. Figure 3)

[3]For MOS ranks were calculated by averaging the repeatedly measured judgements for each sentence per participant. The similarity-difference ratings in part 1 were transformed into ranks for each system by averaging the values of their distance of their two sentences to stimulus T1, as generated in a weighted Euclidean MDS measure.

rank	wE, ind. trns	wE, simlut. trns	sE	MOS
1	T1	T1	T1	T1
2	B1	B1	B1	B1
3	B4	T5	B4	B4
4	T5	B4	T5	T5
5	B3	T2	B3	T4
6	T4	B5	T4	T2
7	T2	T4	T2	B3
8	B2	B3	T3	B5
9	T3	T3	B5	T3
10	B5	B2	B2	B2

Figure 3: Ranks of stimuli, computed as their distance from stimulus T1 in two-dimensional Weighted Euclidean MDS (with individual and simultaneous transformations) and Simple Euclidean MDS, respectively, and ranks from MOS tasks

When comparing the ranks computed from distances generated by Simple Euclidean to ranks as they are established by other MDS models and their correlation with MDS, it is indeed the one that comes closest to duplicating the ranks generated from MOS.

Even though ranks only give a rough approximation, and are very vague, they show that there is a certain consensus across all MDS models as well as MOS, and the rough order is very similar. Although some stimuli vary slightly in their ranks, for each stimulus there is no question whether it is more in the front, the middle, or the back of the field.

Multiple linear regression was then performed to further investigate the nature of the relation between two-dimensional Weighted Euclidean MDS representation and MOS. Since d2 has been shown to have a stronger correlation with MOS than d1, d2 was used as first input variable in blockwise entry of variables in linear regression. d2 accounts for more than 70% of variability in MOS, and together, d1 and d2 account for 95.4%.

The regression was a good fit ($R^2_{adj} = 94\%$), and the overall relationship is significant ($F_{2,7} = 72, p < 0.01$). With other scores held constant, MOS were positively related to dimensions 2 and 1, increasing by 1.387 and 1.212 for every unit, respectively. All effects were significant at $p < 0.01$ Thus we can estimate MOS from our MDS representation as follows:

$$MOS = 5.758 + (1.387 * d2) + (1.212 * d1) \qquad (1)$$

The rms error between measured and predicted MOS is 0.579 per unit. This score is not at all bad; when comparing it to that achieved by [7], we must bear in mind that MOS in this experiment ranged from 1 to 10, as opposed to Hall's range of 1 to 5. Consequentially we are faced with bigger variance, which influences rms.

What becomes obvious in the comparison of stimuli as well as systems is that the systems perceived as most natural and as least natural are the same across test- and analysis methods. However, this is not entirely the case for the intermediate stimuli / systems.

4.3. Comparing direct and aggregated data

In the evaluation of the Blizzard Challenge, judgments are aggregated across listeners as well as across stimuli of one system. In order to avoid including any further noise in the data, only native speakers' judgments of test set A were evaluated by the author, and the number of listeners within each test group and the number of stimuli each listener judged were kept constant. This meant that a lot of listener judgments were lost, particularly because for a large number of speakers the variable of

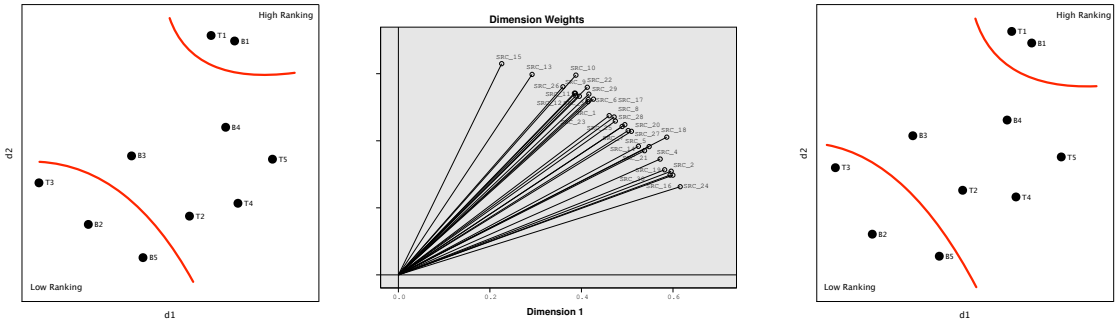

Figure 2: 2-dimensional Weighted Euclidean MDS, ordinal level, untieing ties (left), subject space, showing the relative weights listeners appoint to the dimensions (middle), and 2-dimensional Simple Euclidean MDS, ordinal level, untieing ties, flipped and rotated to align to the axes of Weighted Euclidean MDS representation (right)

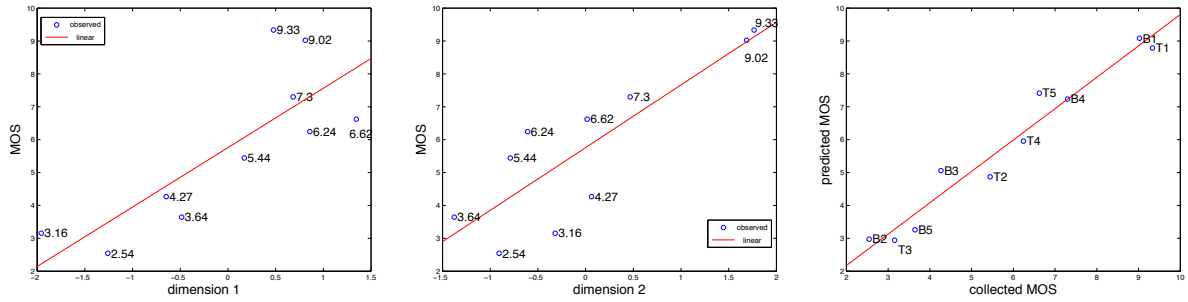

Figure 4: MOS plotted vs dimension 1 (left) and dimension 2 (middle), predicted v. measured MOS(right)

native language had not been defined. However, it seemed the smaller price to pay. Listeners, whose judgments were included in the analysis, were chosen as follows: during the experiment, each listener was appointed to a test group, which determined the subset of stimuli they were presented with. For the analysis, listeners were sorted within the 21 test groups according to their case ids. Then all listeners, who had not completed 21 similar-different judgments were excluded. The numbers of remaining listeners within each group were compared, the lowest of which was 6. The data generated by the first 6 listeners from each group was processed for MDS evaluation, all other data was discarded. *similar* judgments were appointed a value of 0, *different* judgments a value of 1. These values were summed across participants across groups across stimuli for each system. The resulting 21x21 matrix was then used as input into PASW 17.

Two-dimensional ordinal Simple Euclidean distance scaling was performed, untieing ties, resulting in a stress-1 of 0.22, and D.A.F of 0.95. The stress level is above what generally is considered acceptable, but for the sake of ease of comparison with our previous MDS representations we will work with this one, anyway, rather than plotting it in more dimensions.

Interestingly, there is a match between the ranks computed for the systems on the basis of distances derived by MDS from the aggregated Blizzard data, and the ranks computed from the averaged distances for each system's two sentences in Weighted Euclidean MDS on an ordinal level, untieing ties, applying transformations simultaneously, in our smaller-scale experiment. This supports the assumption that aggregated data will indeed provide reliable results that are representative for a larger population. This claim in turn is supported by the fact that if only one stimulus per system is considered (e.g. all B stimuli for a system), the resulting ranks can be very different. This supports our initial hypothesis that averaging over a good stimulus and a bad stimulus of a system could indeed be a good approach towards approximating the results of larger scale evaluations. Further testing on a larger scale will be needed to investigate how reliable that measure is, but these initial findings here are very promising.

5. Discussion

These results show that if a group is homogeneous enough - which a group of native speakers of English seems to be - configurations relying on data that is aggregated across listeners is an acceptable representation of the single listener. If stimuli are chosen appropriately, the aggregation of data generated by a small number can be representative of a larger number. This hypothesis must still be tested more extensively, but if it holds, the amount of data required for evaluation can be drastically minimized. Even though complete data matrices are intuitively more exact, an approximation relying on less data may actually be desirable for many reasons:

> A price is paid for data, not only in financial terms but in wear and tear on the organism at source. A method with too high a channel capacity may, through boredom and fatigue, result in a decrease in information transmitted, through stereotype of behavior. Furthermore, the potential variety of messages from the organism may not be great, in

> which case a more powerful method is inefficient. [...] Ideally a method should be selected which matches the information content in the source but is not such a burden as to generate noise [3, p.51].

So what implications does that have for evaluation projects like the Blizzard Challenge? First of all, these experiments have shown that Simple Euclidean MDS is an appropriate representation for individual listeners' judgement of speech, which is the precondition for data aggregation, which in turn is the precondition for large-scale evaluation. Secondly, the data suggest that MDS offers the same information content as MOS and beyond them, which justifies the elaborate data collection. And thirdly, and probably most exciting, is the suggestion that given a representative subset, smaller-scale experiments can predict larger-scale experiments' results. This initial result justifies more research on how to find representative stimuli, spanning the range of stimuli of a system. If further tests in averaging stimuli of a system prove successful, it should indeed be feasible to reduce the number of participants needed in the MDS part of evaluation. However, in order to do so, the experimental environment will need to be very controlled: to reduce variance of the judgments, participants should all be native speakers and tested in the same, quiet environment, using the same equipment. Recordings of the time elapsed between initial presentation with a stimulus and the point when a decision is entered could be used as a further measure for checking distribution/dispersion of stimuli (cf. [7]) as well as of listeners. MOS, which are collected anyway, can also be used to check for individual's biases, and collectively as a reference frame to check whether the resulting MDS representations are plausible. Alternatively, as the MDS ranks correlate highly with the MOS, the MOS tests could be dropped in favour of MDS tests which provide more informative results in terms of rankings and spacial layout.

The stimuli chosen as representatives of a system can be derived in a test series prior to the main evaluation: Similarity-difference tests are conducted in the same manner as used in the experiment described in this paper, using a natural stimulus and a few test sentences from a system. MDS is performed and the sentence with the biggest and that with the smallest distance from the natural system is selected. This is done for all systems to be tested, and the thus selected stimuli are then used in large scale evaluation. This part in itself is a fairly expensive again, but the stimuli thus picked will remain representative of a system, and this part of testing will not have to be repeated, unless changes are made to the system. Hence, in future years, it will be less costly to include more systems from previous years into the Blizzard Challenge for the sake of anchoring. This small step towards a benchmark is a great improvement in subjective evaluation of synthesized speech.

Acknowledgements

The authors would like to thank The Blizzard Challenge organisers and participants for providing the data used for this work.

References

[1] Alvarez, Y. and Huckvale, M. [2002], The reliability of the ITU-T P. 85 standard for the evaluation of text-to-speech systems, *in* 'Seventh International Conference on Spoken Language Processing', ISCA.

[2] Borg, I. and Groenen, P. [2005], *Modern multidimensional scaling: Theory and applications*, Springer Verlag.

[3] Coombs, C. H. [1964], *A theory of data*, Wiley.

[4] Coxon, A., Jackson, J., Davies, P., Smith, H., Sachs, L. and Schmee, J. [1982], *User's guide to multidimensional scaling*, Heineman Education books.

[5] Coxon, A. M. [2003], Multidimensional scaling, *in* M. Lewis-Beck, A. Bryman and T. Liao, eds, 'The Sage encyclopedia of social science research methods', Sage Publications, Inc.

[6] Fraser, M. and King, S. [2007], The blizzard challenge 2007, *in* 'Proc. Blizzard Workshop (in Proc. SSW6)'.

[7] Hall, J. [2001], 'Application of multidimensional scaling to subjective evaluation of coded speech', *The Journal of the Acoustical Society of America* **110**, 2167.

[8] Hirst, D., Rilliard, A. and Auberge, V. [1998], Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis, *in* 'The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis', ISCA.

[9] Holmes, J. [2001], *Speech synthesis and recognition*, CRC.

[10] Jurafsky, D. and Martin, J. [2008], *Speech and language processing*, Prentice Hall.

[11] Karaiskos, V., King, S., Clark, R. and Mayo, C. [2008], 'The Blizzard Challenge 2008'.
URL: *http://festvox.org/blizzard/bc2008/summary_Blizzard2008.pdf*

[12] Mayo, C., Clark, R. and King, S. [2005], Multidimensional scaling of listener responses to synthetic speech, *in* 'Ninth European Conference on Speech Communication and Technology', ISCA.

[13] Meulman, J., Heiser, W. and SPSS, I. [2001], 'Categories 11.0', *Chicago: SPSS Inc* .
URL: *http://www.courses.rochester.edu/SPSSDocs/SPSS%20Categorie%2011.0*

[14] Podsiadlo, M. [2007], Large scale speech synthesis evaluation, Master's thesis, University of Edinburgh.

[15] Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A., Swerts, M. and Terken, J. [1998], Evaluation of speech synthesis systems for Dutch in telecommunication applications, *in* 'The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis', ISCA.

[16] Taylor, P. [2009], *Text-to-Speech Synthesis*, Cambridge University Press.

[17] Vainio, M., Jarvikivi, J., Werner, S., Volk, N. and Valikangas, J. [2002], Effect of prosodic naturalness on segmental acceptability in synthetic speech, *in* 'Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on', pp. 143–146.